REPORT DOCUMENTATION PAGE

1

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	New Reprint	-

4. TITLE AND SUBTITLE
Sparse-Representation-Based Classification withStructure-Preserving Dimension Reduction

5a. CONTRACT NUMBER
W911NF-12-1-0378

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Jin Xu,, Guang Yang,, Yafeng Yin,, Hong Man,, Haibo He,

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Rhode Island
Sponsored Projects
70 Lower College Road, Suite II
Kingston, RI 02881 -1967

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
61817-CS.21

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Sparse-representation-based classification (SRC), which classifies data based on the sparse reconstruction error, has been a new technique in pattern recognition. However, the computation cost for sparse coding is heavy in real applications. In this paper, various dimension reduction methods are studied in the context of

15. SUBJECT TERMS
Sparse representation (coding) , ┐Classification, ┐ Feature extraction, ┐ Feature selection ,┐Dimension reduction, ┐ Structure preserving

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Haibo He
UU	UU	UU	UU		19b. TELEPHONE NUMBER 401-874-5844

Report Title

Sparse-Representation-Based Classification withStructure-Preserving Dimension Reduction

ABSTRACT

Sparse-representation-based classification (SRC), which classifies data based on the sparse reconstruction error, has been a new technique in pattern recognition. However, the computation cost for sparse coding is heavy in real applications. In this paper, various dimension reduction methods are studied in the context of SRC to improve classification accuracy as well as reduce computational cost. A feature extraction method, i.e., principal component analysis, and feature selection methods, i.e., Laplacian score and Pearson correlation coefficient, are applied to the data preparation step to preserve the structure of data in the lower-dimensional space. Classification performance of SRC with structure-preserving dimension reduction (SRC–SPDR) is compared to classical classifiers such as k-nearest neighbors and support vector machines. Experimental tests with the UCI and face data sets demonstrate that SRC–SPDR is effective with relatively low computation cost

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

Continuation for Block 13

ARO Report Number 61817.21-CS
Sparse-Representation-Based Classification wit...

Block 13: Supplementary Note

Approved for public release; distribution is unlimited.

Cogn Comput
DOI 10.1007/s12559-014-9252-5

Sparse-Representation-Based Classification with Structure-Preserving Dimension Reduction

Jin Xu · Guang Yang · Yafeng Yin · Hong Man · Haibo He

Received: 14 September 2012 / Accepted: 18 February 2014

Abstract Sparse-representation-based classification (SRC), which classifies data based on the sparse reconstruction error, has been a new technique in pattern recognition. However, the computation cost for sparse coding is heavy in real applications. In this paper, various dimension reduction methods are studied in the context of SRC to improve classification accuracy as well as reduce computational cost. A feature extraction method, i.e., principal component analysis, and feature selection methods, i.e., Laplacian score and Pearson correlation coefficient, are applied to the data preparation step to preserve the structure of data in the lower-dimensional space. Classification performance of SRC with structure-preserving dimension reduction (SRC–SPDR) is compared to classical classifiers such as k-nearest neighbors and support vector machines. Experimental tests with the UCI and face data sets demonstrate that SRC–SPDR is effective with relatively low computation cost

Keywords Sparse representation (coding) · Classification · Feature extraction · Feature selection · Dimension reduction · Structure preserving

J. Xu
Operations and Information Management Department, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: jinxu@wharton.upenn.edu

G. Yang · Y. Yin · H. Man
Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA
e-mail: gyang1@stevens.edu

Y. Yin
e-mail: yyin1@stevens.edu

H. Man
e-mail: Hong.Man@stevens.edu

H. He (✉)
Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881, USA
e-mail: he@ele.uri.edu

Introduction

In recent years, sparse representation (or sparse coding) has received a lot of attentions. The key idea is to search for the least number of basis vectors (or atoms) in a dictionary $\mathbf{A} \in \mathbb{R}^{m \times n}$ to characterize a signal $\mathbf{y} \in \mathbb{R}^m$ ($\mathbf{A}$ has n atoms and each atom is a vector with m elements). Therefore, the signal can be represented as the sparse vectors $\mathbf{x} \in \mathbb{R}^n$ based on atoms. The atoms in $\mathbb{R}^m$ are the column vectors in $\mathbf{A}$. Sparse representation improves performance in a number of applications [64], such as coding [42], classification [62], image denoising [13], smart radio [30, 31], dimension reduction [18, 60] and so on.

Sparse coding has extensive connections to biological-inspired and cognitive approaches. In [42], the properties of the primary visual cortex are used to interpret sparse linear codes. In the research of V1 simple cell receptive fields [70], the sparse coding is trained using biologically realistic plasticity rules. In [16], sparse coding is used to explain brain function in primate cortex. In [49], extracting covariance patterns based on sparse coding gives a promising direction in cognitive brain region identification. The large-scale brain modeling is a promising direction in cognitive science. In [43], the application of sparse coding in associative memory pattern has been pointed out, which could contribute to the complicated brain modeling. Recently, structured sparse coding

Published online: 13 March 2014

has been proposed based on neocortical representations [23].

In this paper, we focused on classification based on sparse representation in low dimension. The work is partially inspired by a sparse-representation-based classification (SRC) method recently proposed in [56], which searches for the training samples producing the minimum reconstruction error of testing data. Results reported in [56] were very promising and competitive to those from traditional classification methods, such as support vector machine (SVM) and k-nearest neighbors (KNN).

It is well known that sparse representation methods are computationally intensive. The number and dimension of atoms in a dictionary affect computation cost significantly. In the community, there are three techniques to reduce the computational complexity of sparse coding:

- Structure-preserving dimension reduction (SPDR): The purpose is to reduce redundancy as well as retain structure in the data preparation process. Many researchers have devoted their work to achieve this goal [4, 20, 24, 45, 66]. Various classic dimension reduction methods have been applied to sparse coding. Sparse latent semantic analysis (sparse LSA) was proposed in [8], the sparsity constraint via the ℓ_1 regularization was added in the formulation of the LSA, which is a popular unsupervised dimension reduction tool. Experimental results show that sparse LSA could be effective at reducing the cost of projection computation and memory. The multi-label sparse coding framework with feature extraction [52] was applied to automatic image annotation, and comparisons with state-of-the-art algorithms demonstrated its efficiency. In our previous work [59], we significantly extended definitions for the sparse representation method and investigate its analytical characteristics as well as empirical results.
- Dictionary construction: The key to successful sparse coding lies in the dictionary. There are two main approaches to constructing a dictionary: analytic design and dictionary learning. Analytic design establishes proper atoms from abstract function spaces [34] or pre-constructed dictionaries, such as wavelets [40] and contourlets [5]. In dictionary learning [36], various technologies such as regularization and clustering are applied on training data to build dictionary. In [14], the least-square error was utilized via the method of optimal directions (MOD) to train dictionary. Online dictionary learning [39] used stochastic approximations to update dictionary with a large data set. Laplacian score dictionary (LSD) [58], which is based on the geometric local structure of training data, selected the atoms for the dictionary.
- Efficient optimization algorithm: Different optimization methods are embedded the sparse coding process to improve computational efficiency. A convex version of sparse coding was proposed in [3], a regularization function via compositional norms was implemented in convex coding, and boosting-style algorithm was derived. Experimental results in the image denoising task showed the advantages of the boosted coding algorithm. In efficient sparse coding algorithms [38], ℓ_1 regularized and ℓ_2 constrained least-squares problem was solved iteratively, and its applications on image process showed the significant acceleration for sparse coding. In [22], a nonlinear feed-forward predictor was trained to produce the sparse code, and the proposed method required 10 times less computation cost than previous competitors.

In this paper, we present a combined SRC and SPDR framework. Dimension reduction can effectively reduce the computation cost and extract useful structural information. It can also contribute to improved performance recognition tasks: (i) Discriminative learning for dimensionality reduction was proposed in [37]. A supervised form of latent dirichlet allocation (LDA) was derived. The class label information was incorporated into LDA, which enabled the discriminative application of LDA. (ii) A five-step procedure, which increased different dimension reduction methods with classification, is proposed in [11]. In particular, partial least squares (PLS), sliced inverse regression (SIR) and principal component analysis (PCA) were compared in terms of classification performance with gene expression data sets.

Similarly in our work, four dimension reduction methods, i.e., PCA, Laplacian score (abbreviated as LAP), Pearson correlation coefficient (abbreviated as COR) and minimum-redundancy maximum-relevancy (abbreviated as mRMR) [45] are studied in the SRC framework, and extensive experiments in comparison with other classic classifiers (SVM and KNN) are carried out.

The contributions of this paper can be summarized as follows:

- A comprehensive study of various SPDR methods in sparse representation is presented. In particular, the performance of feature extraction and feature selection methods are examined.
- The proposed methods are successfully applied to both the UCI data sets and face image data sets. While most sparse coding work has concentrated on natural signal and image data sets, very few have applied sparse coding to the feature space data sets (UCI data sets).
- Very competitive classification results are obtained on both UCI data sets and face data set, providing new insight to the capabilities of sparse representation.

The rest of the paper is organized as follows: Sect. 2 reviews related SPDR methods. Section 3 presents our proposed SRC based on SPDR. Section 4 presents experimental results of this framework on the UCI data sets and face data sets. Finally, Sect. 5 gives conclusions and discusses future works.

Classic Dimension Reduction Methods with Structure Preserving

Dimension reduction is an important method in knowledge discovery [21] and machine learning [57]. Structure-preserving [47] constraints can improve dimension reduction. Normally, certain compact coordinates are obtained by dimension reduction methods to preserve special properties of the input data. The distances properties of data points were preserved by the multidimensional scaling [10]. The local geometry of the data set was studied by nonlinear manifold learning [54]. Eigenvector-based multivariate analysis [51] revealed the internal structure in terms of variance.

Many sparse-representation-based dimension reduction algorithms have been developed extensively recently, including elastic net [68], sure independence screening [15] and the Dantzig selector [7]. These researches are normally focused on reducing the number of atoms for sparse representation, such as setting certain sparse coefficients to zeros. For example, in group structure sparsity [32] and tree structure sparsity [35], the sparse coefficients were modified based on this prior information. However, there is little work on exploring the relationship between lower-dimension data sets and sparse representation.

There are two categories for dimension reduction, feature extraction (such as PCA) and feature selection. PCA is a linear transformation that best represents the data in the least-squares sense. Any signal can be coarsely reconstructed as a linear combination of principal components. Sparse PCA [69] was proposed based on lasso constraints with the result of sparse loading. In terms of feature selection [67], it focused on searching for a subset of features from the original feature sets. Some feature selection methods [48, 61] combined with sparse representation have been shown to be effectiveness.

A huge volume of literature is devoted to projecting high-dimensional data to a lower dimensional space through various methods, such as: locally linear embedding, linear discriminant analysis, PCA, LAP [29] and COR [24]. We just choose four of them to combine with SRC, following the previous work [56] that SRC is not sensitive to a particular projection method.

PCA Criteria and Eigenface

PCA was first used on face recognition by Turk and Pentland [51], which is now known as eigenfaces. Given a training set of face images $I_1, I_2, \ldots, I_n$, the first step is to represent each image I_i with a vector Γ_i, and then subtract the average face $a = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i$ from the training face image vector $\Phi_i = \Gamma_i - a$.

Next, eigenvalues and eigenvectors of the covariance matrix C can be obtained.

$$C = \frac{1}{n}\sum_{i=1}^{n}\Phi_n\Phi_n^T \tag{1}$$

Typically, only the n most significant eigenvalues and their corresponding eigenvectors are calculated. The resulting eigenvectors are the eigenfaces. Each test image will be projected onto these eigenvectors, and the coefficient vectors are then used in the classification.

PCA is a popular dimension reduction method, which projects the data in the direction of maximal variances to obtain the minimized reconstruction error. Normally, it is a linear data transformation to preserve the global structure, but kernel-based PCA could be applied to nonlinear problems. A PCA based method has successfully applied in identification of human population structure [44].

Laplacian Criteria for Dimension Reduction

The Laplacian method preserves local geometrical structures without the data labels. LAP [29] is a new feature selection method based on Laplacian eigenmaps and locality-preserving projection. The score evaluates the feature's importance according to its locality-preserving ability.

For the data $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$ with feature set $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m]$, assume V_r is the LAP for the rth feature $\mathbf{f}_r$, the LAP is calculated as:

1. It first constructs a nearest neighbor graph G with different data nodes ($\mathbf{y}_i$ and $\mathbf{y}_j$, $i,j = 1, \ldots, n$) in data sets. $S_{ij} = e^{-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{t}}$ represents the score between data $\mathbf{y}_i$ and $\mathbf{y}_j$, where t is a suitable constant.
2. Then S_r can be defined as

$$V_r = \frac{\tilde{\mathbf{f}}_r^T L \tilde{\mathbf{f}}_r}{\tilde{\mathbf{f}}_r^T D \tilde{\mathbf{f}}_r} \tag{2}$$

where $D = diag(S\mathbf{1}), \mathbf{1} = [1, \ldots, 1]^T, L = D - S$, and $\tilde{\mathbf{f}}_r$ is a kind of normalization via:

$$\tilde{\mathbf{f}_r} = \mathbf{f}_r - \frac{\mathbf{f}_r^T D \mathbf{f}_r}{\mathbf{1}^T D \mathbf{1}} \tag{3}$$

In LAP, the nearest neighbor graph based on the data points is established, and the local structure is evaluated by the weights between nodes. Therefore, the structure in the graph preserves the discriminate features in the feature space.

Pearson Correlation Coefficient Criteria and mRMR

Pearson correlation coefficient is based on the covariance matrix and can select the feature variable with target labels [24]. Normally, it is a supervised feature selection method. The COR between two different variables is:

$$P(\alpha_i, \alpha_j) = \frac{\mathrm{cov}(\alpha_\mathrm{i}, \alpha_\mathrm{j})}{\sqrt{\mathrm{var}(\alpha_i) \times \mathrm{var}(\alpha_j)}} \tag{4}$$

According to the max-dependency and min-redundancy [45] concepts, the feature selection process can combine the dependency and redundancy criteria together. In this work, these criteria are used in the COR. First, max-relevance criteria are applied from different features ($\mathbf{f}_i \in \mathbf{F}$) to the target c to get the most relevant feature.

$$\max D(\mathbf{F}, \mathbf{c}), \quad D = \frac{1}{|\mathbf{F}|} \sum_{\mathbf{f}_i \in \mathbf{F}} P(\mathbf{f}_i, \mathbf{c}) \tag{5}$$

Then for min-redundancy criteria, the selected feature is:

$$\min R(\mathbf{F}), \quad R = \frac{1}{|\mathbf{F}|^2} \sum_{\mathbf{f}_i, \mathbf{f}_j \in \mathbf{F}} P(\mathbf{f}_i, \mathbf{f}_j) \tag{6}$$

In order to combine max-relevance and min-redundancy, an operator $\Phi(D, R)$ is defined. It is a simple form of optimization of D and R.

$$\max \Phi(D, R), \quad \Phi = D - R \tag{7}$$

In the process of incremental feature selection [45], suppose we have chosen $\{m-1\}$ features with the feature set $\mathbf{F}_{m-1}$. In the m step, it selects the mth feature from feature set $\{\mathbf{F} - \mathbf{F}_{m-1}\}$. This can be operated from $\Phi(D, R)$. The criteria are:

$$\max_{\mathbf{f}_j \in \{\mathbf{F} - \mathbf{F}_{m-1}\}} \left[P(\mathbf{f}_j, c) - \frac{1}{m-1} \sum_{\mathbf{f}_i \in \mathbf{F}_{m-1}} R(\mathbf{f}_j, \mathbf{f}_i) \right] \tag{8}$$

Our method constructs sub-feature space with the features most connected with the class (category) center while filtering redundant features, which are criteria of dependence structure between the features and class centers. For the mRMR method, it uses mutual information to build the relations between the features, which is popular and robust in many applications, the detailed settings are described in [45].

SRC Based on Dimension Reduction

In sparse representation, assume a dictionary with a set of training data vectors (or atoms) $\mathbf{A} = [\mathbf{a}_1^1, \ldots, \mathbf{a}_1^{n_1}, \ldots, \mathbf{a}_c^1, \ldots, \mathbf{a}_c^{n_c}]$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$, c is class label for each atom, n_i is the number of atoms associated with the category i. Then a new test data vector $\mathbf{y}$ is represented in the form:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \in \mathbb{R}^m \tag{9}$$

where $\mathbf{x} = [0, \ldots, 0, \alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,n_i}, 0, \ldots, 0]^T \in \mathbb{R}^n$ is the sparse vector (coefficients). In order to calculate $\mathbf{x}$, we use the ℓ_1-regularized least-squares method [33, 50] defined as:

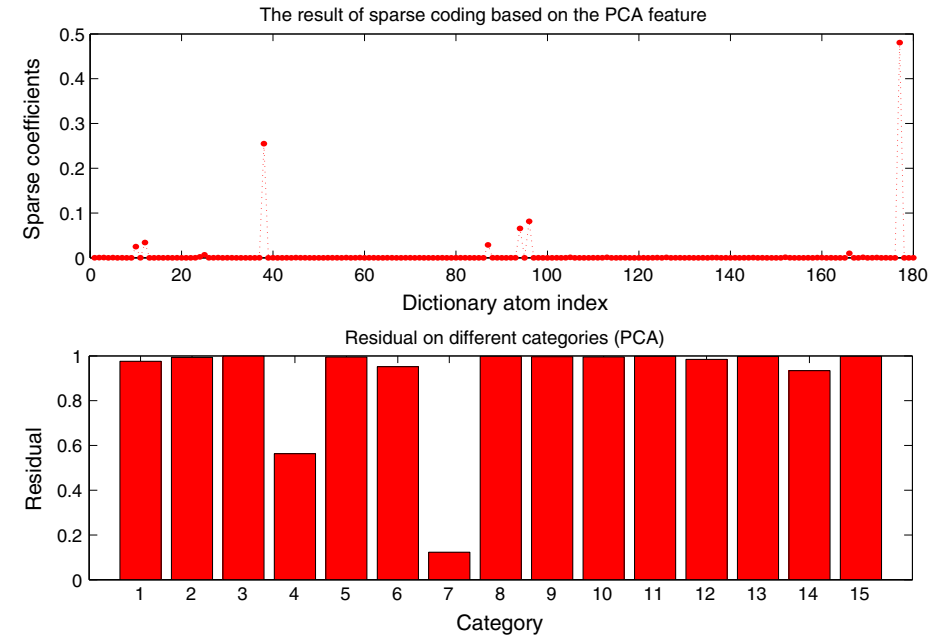
$$\widehat{\mathbf{x}} = \arg\min \{ \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \} \tag{10}$$

In [56], SRC utilizes the representation residual to predict class labels for test samples. In particular, a characteristic function ($\delta_i : \mathbb{R}^n \to \mathbb{R}^n$) is defined for each category i, which chooses the sparse coefficients via the category. And the classification is based on:

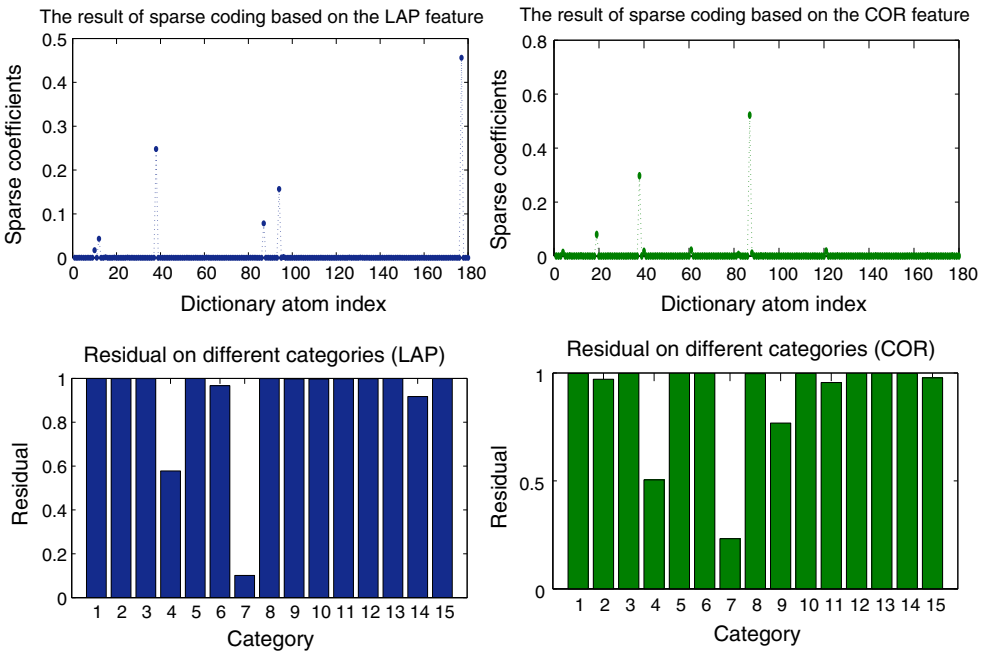
$$\mathrm{label}(\mathbf{y}) = \arg\min r_i(\mathbf{y}), \quad r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\delta_i(\mathbf{x})\|^2 \tag{11}$$

In their work, some related dimension reduction methods with SRC are combined to show that SRC is robust with low-dimensional features from images. In particular, random face, downsampling face and Fisher face are used as low-dimensional features for SRC. The paper claims that the choice of dimension reduction methods does not significantly impact SRC performance, and sufficient dimensionality (such as dimension 100 for face data) of the data is more important for SRC.

We follow this direction to propose the SPDR method, which we apply on the UCI and image data sets. In details, a dimension reduction projection $P_{\mathcal{S}(\mathcal{B})}$ is applied to the input data $\mathbf{Y}$ and all the atoms in $\mathbf{A}$, where subspaces are denoted by $\mathcal{S}$, and $\mathcal{S}(B)$ means the subspace spanned by matrix B, the dimension of the data would be changed from dimension m to dimension d ($m > d$). In our work, the matrix B is obtained from PCA, LAP, COR and mRMR. Algorithm 1 shows in detail the procedure of the SRC method with SPDR. In classification, a function ψ is built from a training set $(\mathbf{Y}_i, c_i)$, $i = 1, \ldots, n$. The goal of



(a) PCA feature SRC



(b) LAP feature SRC **(c)** COR feature SRC

Fig. 1 Sparse-representation-based classifier is applied to the "Libras Movement" data set with 3 different feature selection methods. For each case, the *upper figure* shows sparse coefficients based on the corresponding dictionaries (the *dots* denote the non-zero coefficients), and the *lower figure* shows representation residuals $r_i(y)$ on different categories

dimension reduction for classification is to retrieve subspaces [9, 53] which are the most relevant to the classification, noted as subspace $\mathcal{S}(B)$ such that:

$$\varphi(\mathbf{Y}) = \varphi(P_{\mathcal{S}(\mathcal{B})}\mathbf{Y}) \tag{12}$$

the decision rule φ established from the projected data $P_{\mathcal{S}(\mathcal{B})}\mathbf{Y}$ should be the same as that established from the original data $\mathbf{Y}$.

A case study of this algorithm is shown in Fig. 1. The algorithm is run over the "Libras Movement" data set from the UCI source [17]. The dimension of data is reduced from original 90 to 20 via PCA, LAP and COR, respectively. SRC is then applied to a test vector $\mathbf{y}$ based on a dictionary containing 180 training vectors. The SRC on test data is shown in Fig. 1. In particular, the sparse coefficients based on the dictionary and the corresponding representation residuals on different classes are exhibited. We can observe that:

Algorithm 1 SRC-SPDR

1: **Input:** a set of training data $\mathbf{A} \in \mathbb{R}^{m \times n}$ with c classes, test data $\mathbf{y} \in \mathbb{R}^m$, a dimension reduction matrix B
2: Compute new data $\tilde{\mathbf{A}} = P_{\mathcal{S}(B)}A$ and $\tilde{\mathbf{y}} = P_{\mathcal{S}(B)}\mathbf{y}$
3: Solve ℓ_1-regularized least squares problem:
$\widehat{\mathbf{x}} = arg\min\{\|\tilde{\mathbf{y}} - \tilde{\mathbf{A}}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1\}$
4: Compute the residuals:
$r_i(\mathbf{y}) = \|\tilde{\mathbf{y}} - \tilde{\mathbf{A}}\delta_i(\mathbf{x})\|^2 \quad for \quad i = 1, \cdots, c$
5: **Output:** $label(\mathbf{y}) = arg\min r_i(\mathbf{y})$

- The classification performance is the same in each of these three cases, and the dimension-reduced data are sufficient for SRC to make judgments.
- The coefficients from LAP and COR are sparser than those from PCA.
- Category 4 has the second smallest residual, so it provides an indication of the similarity between category 4 and category 7. There may be potential for SRC to cluster similar categories.
- The results from PCA and LAP are similar, partially because they are both unsupervised feature selection methods. COR, which is a supervised feature selection method, produces results that are quite different from the others.

Complexity Discussion

Suppose we have a signal $f \in \mathbb{R}^N$, it can be decomposed into an orthonormal basis $\boldsymbol{\Psi}$ and a coefficient vector $\boldsymbol{x}$, which can be written in the following way:

$$\boldsymbol{f} = \sum_{i=1}^{N} x_i \psi_i = \boldsymbol{\Psi x} \tag{13}$$

The vector $\boldsymbol{x}$ is $N \times 1$ dimension, and orthobasis $\boldsymbol{\Psi}$ is in $\mathbb{R}^{N \times N}$. $\boldsymbol{x}$ is called *S-sparse* if it has at most S non-zero elements. The signal $\boldsymbol{f}$ is composed of a best subset of S columns that span in an orthogonal basis of size $N \times N$. It implies that the choices of finding out this particular subset is $\binom{N}{S}$.

To reconstruct the signal $\boldsymbol{f}$ from a linear combination of vectors, we would like to constrain the error to be smaller than a fixed approximation error as well as only keep the least number of vectors that are orthogonal with each other. However it is an NP-hard problem [2]. In past literature, before the restricted isometry property (RIP) condition was discovered, the matching pursuit method was developed to solve ℓ_0 norm minimization problems. Here ℓ_0 norm is a pseudo norm and defined as $\|\boldsymbol{x}\|_{\ell_0} \triangleq \#\{i \text{ s.t. x[i]} \neq 0\}$.

In the compressive sensing problems, a measurement space is introduced where there is an observation $\boldsymbol{y} \in \mathbb{R}^m$, and $\boldsymbol{y}$ is obtained by a random sampling matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m \times N}$.

$$\boldsymbol{y} = \boldsymbol{\Phi\Psi x} \tag{14}$$

Therefore, the orthogonal transforming space $\boldsymbol{\Psi}$ is measured by the sampling matrix $\boldsymbol{\Phi}$. Furthermore, there are S vectors selected by the sparse vector x from the matrix $\boldsymbol{\Phi\Psi}$, making up an orthogonal subspace Σ_S, and hence, their linear combinations are the observation of the original vector. From the sampled observation, $\boldsymbol{x}$, the random sensing matrix $\boldsymbol{\Phi}$ can be found under the fixed space $\boldsymbol{\Psi}$, and thereby recover the original signal $\boldsymbol{f}$.

By giving theoretical proof in [6], that the RIP condition holds, ℓ_1 norm minimization can reconstruct the signal as well as the ℓ_0 norm with overwhelmingly high probability if $m = \mathcal{O}(S\log(N/S))$. Furthermore, ℓ_1 norm is a convex optimization, and now can be solved via LASSO regression which is much more tractable than computing the ℓ_0 norm. Lasso has relatively low polynomial computational cost of $\mathcal{O}(m^2N)$ time [41].

Using the sparse representation to reconstruct a signal relies on the same computation framework as the compressive sensing, in the way of sparsely selecting a subset of atoms that are linearly independent with each other, from the over-complete dictionary $A \in \mathbb{R}^{M \times N}(M << N)$.

In order to reduce the expensive computation cost during optimization, we propose a hierarchical sparse coding framework, which has a dictionary with fewer dimensions but without compromising the performance in the multi-label classification task. On one hand, the dimension reduction before sparse coding classification has an overhead computation cost. On the other hand, it reduces the cost in the sparse coding stage. The dimension is reduced in the space of $\mathbb{R}^m$, by selecting certain higher scored elements along the columns of a dictionary. However it has a power to cost in the computation of Lasso optimization $\mathcal{O}(m^2N)$. In our work, we have experimented with reducing m to different levels, not only to lessen the computational cost. In doing so, we discovered that with only a very small number of features, we can preserve the structure of the

dictionary, and thereby keeping the classification performance competitive.

Experimental Results

In this section, we present experimental results on different data sets to study the effectiveness of SRC with dimension reduction. The experiments are conducted on the UCI data sets [17] and the extended Yale face database B [19].

Experimental Setup

In these experiments, we first apply three dimension reduction methods to transform the data to a lower dimensional space. As mentioned before, PCA is a linear data transformation, LAP is an unsupervised feature selection method and COR and mRMR are supervised feature selection methods. Then, we use three classification methods: SRC, SVM and KNN, to show the classification accuracy. SVM is an effective and popular classifier [46], which uses kernel methods to construct class boundaries in higher dimensional space. KNN is a classic classifier and has achieved good performance in recent studies [12].

Each data set is randomly partitioned to training and testing sets at a 1:1 ratio. Each experiment is carried out five times, and the final results are averaged. In the SRC method, the entire training set are included in the dictionary, which is the same setting as the work reported in [56, 63].

In these experiments, the sparse coding software is l_1-ls package [33] from Stanford university, SVM and KNN classifiers are from the Java toolbox [55], and parameters of the tools are set to the default. For SVM and KNN, the parameters are chosen based on the performance over the test set. In detail with SVM, we used three kernels (linear kernel, polynomial kernel and radial basis function kernel), and the kernel parameters are 0.5 and 0.05. There are six outputs for SVM testing. In the KNN side, the number of neighbors "k" is set as 1 or 5, the distances we have used are L1 distance, L2 distance and cosine distance. Therefore, we also get six results for KNN over the test set. The final parameters are those yielded higher average performance over the test set. In Fig. 2, we have shown a case for data glass with PCA dimension reduction. The three best SVM and KNN results are listed.

Experiments on UCI Data Sets

Our experiments cover five benchmark UCI data sets [17]. Due to the difficulty of multi-category classification problems, most selected data sets are multi-category data sets (except "Anneal"). Table 1 shows the detailed information of the experimental data sets.

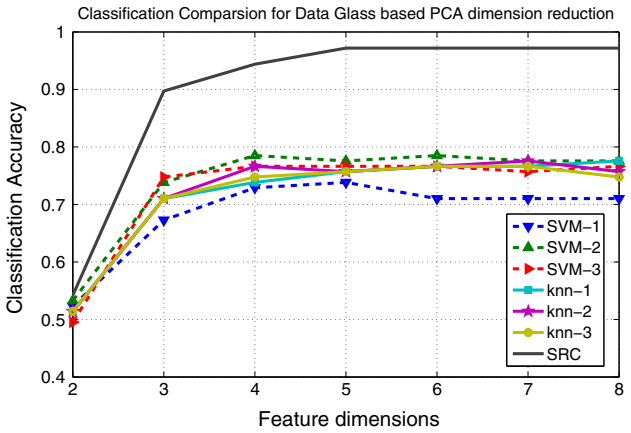


Fig. 2 SVM and KNN with different parameters, 3 best SVM and 3 best KNN performances based on different parameters are shown. In the final comparison, SVM-2 (with polynomial kernel (0.5)) and knn-2 (with 5 neighbors and cosine distance) are chosen

Table 1 UCI data sets

Name	Feature number	Total size	Test size	Class
Wine	13	178	89	3
Glass	10	214	107	7
Libras Movement	90	360	180	15
Wine Quality	11	4,898	2,449	6
Anneal[a]	11	798	399	2

[a] The missing feature has been removed

In these experiments, the dimension of the data is changed from small to large to evaluate the effect of number of features on classification accuracy. Then SRC, SVM and KNN are applied to the lower dimensional data to obtain the classification accuracies. The details are shown in following figures. Figure 3 shows the results for data set "Wine." With PCA and COR, SRC and SVM have higher accuracies in higher dimensions. With LAP, KNN performs slightly better than SVM. In mRMR results, KNN has a higher accuracy than SVM and SRC.

The results for the data set "Glass" are shown in Fig. 4. SRC produces consistently higher accuracies in all three dimension reduction cases. SRC with PCA tends to be stable from dimension 3. SRC with COR reaches more than 95 % from dimension 5 and up, which is much higher than the results of SVM with COR and KNN with COR. For mRMR method, SRC's performance improves dramatically with over 4 features. On this particular data set, our results can be compared with the results presented in [1]. In their work, "Boost-NN", "Allwein" and "Naive k-NN" were applied on the whole data set with size of 214 for training, and the achieved classification rates were 75.6,

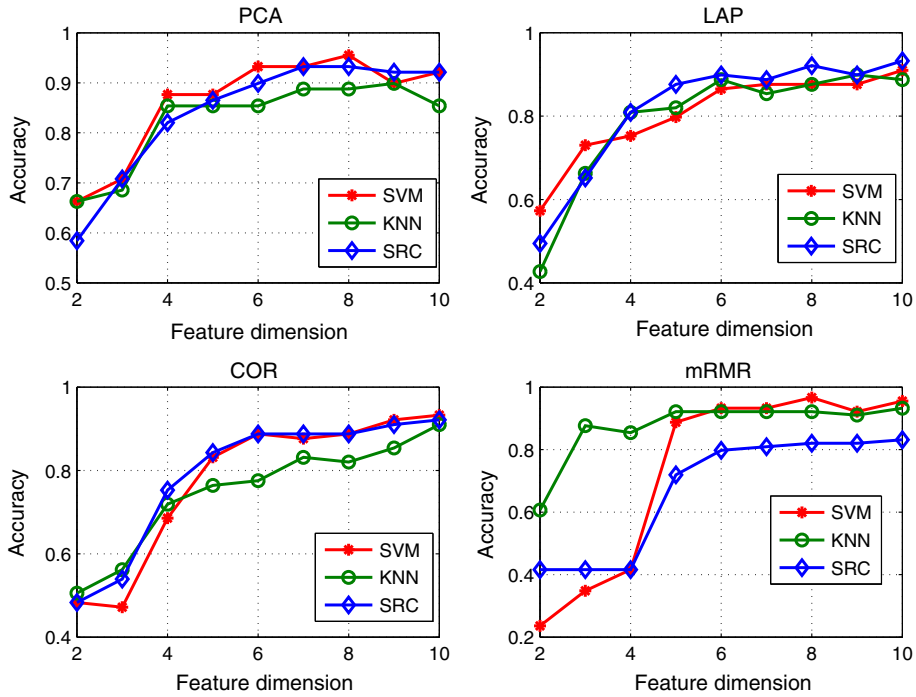


Fig. 3 Classification result for data Wine

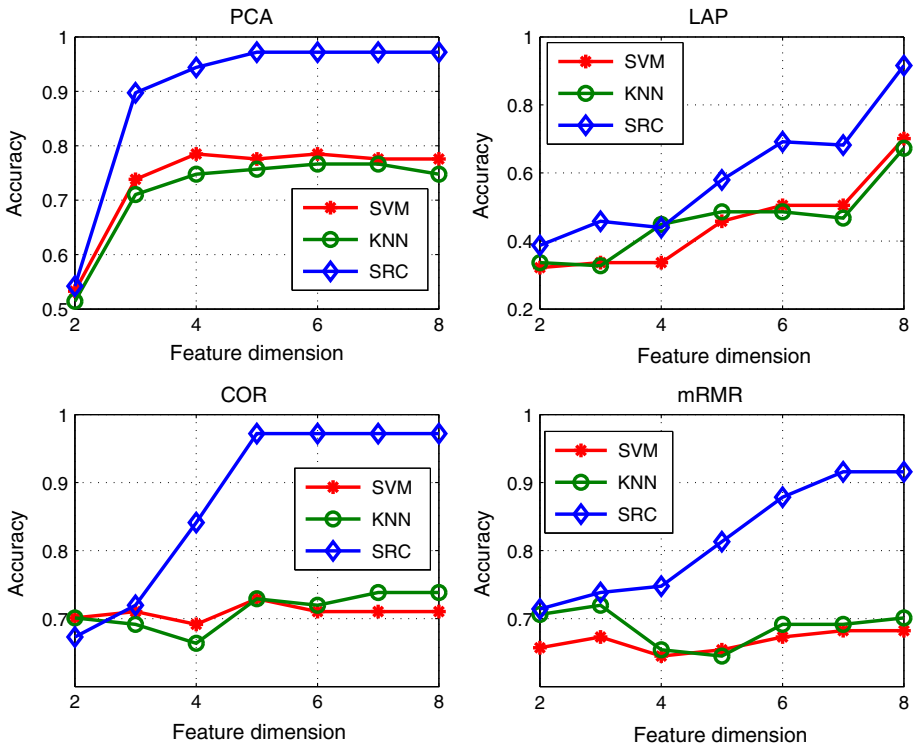


Fig. 4 Classification result for data Glass

74.8 and 73.2 %, respectively. These results are similar to our SVM and KNN results with PCA or COR. However, in our case, a smaller dimension and half of data set are used in training. SRC with PCA or COR clearly results in better performance on these data sets.

In Fig. 5, SRC results show the obvious higher accuracies on the "Libras Movement" data set. In PCA subfigure, when the accuracies of SVM and KNN deteriorate, SRC's accuracy remains at its original level, demonstrating the ability of SRC to deal with noisy data.

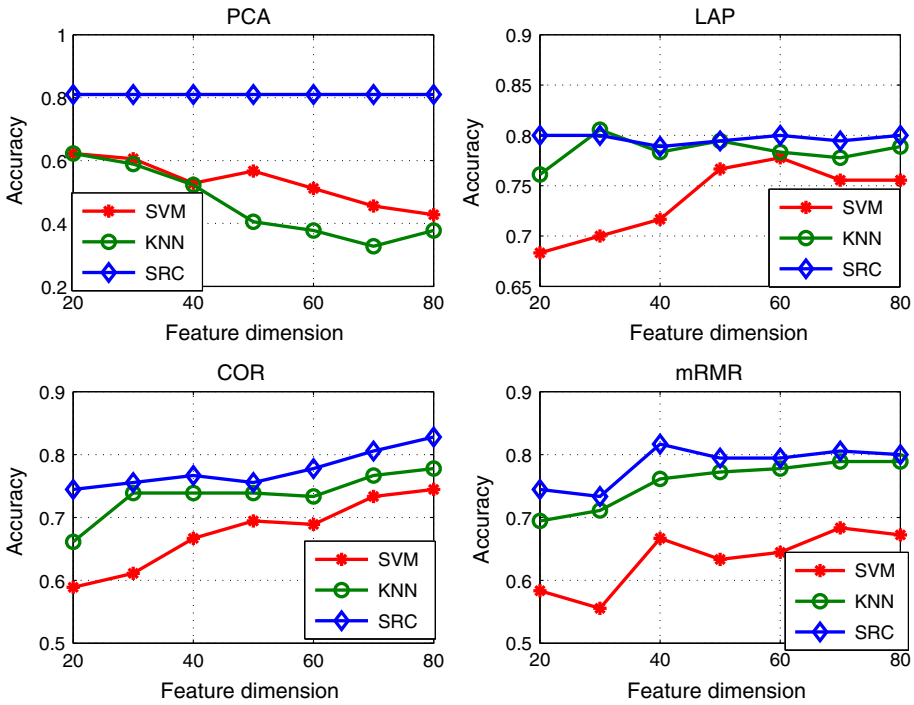


Fig. 5 Classification result for data Libras Movement

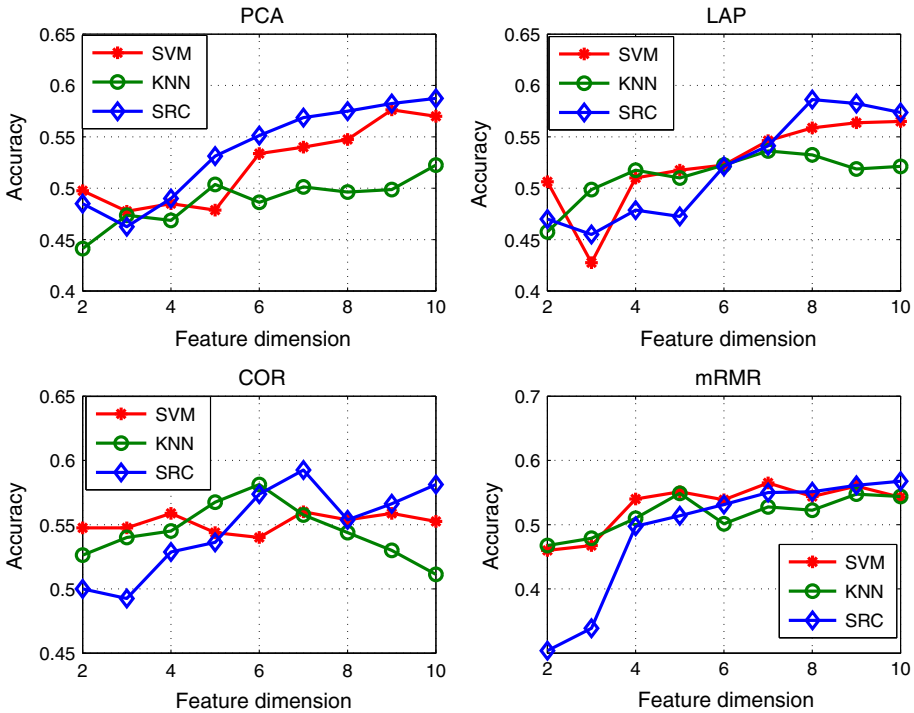


Fig. 6 Classification result for data Wine Quality

For the data set "Wine Quality" in Fig. 6, SRC shows similar performance to SVM and KNN in PCA and LAP features. The accuracies increase steadily as a function of increasing feature dimensions. SRC with PCA tends to have higher accuracy in higher dimensions (from dimension 6 and up). In the case of COR, SRC with COR has the worst performance, which needs to be further investigated. For mRMR, SRC has a stable performance from dimension of 6.

"Anneal" in Fig. 7 is the only binary data set in our experiment. SRC performs better than other classifiers in the PCA case, and SRC shows similar results with KNN in

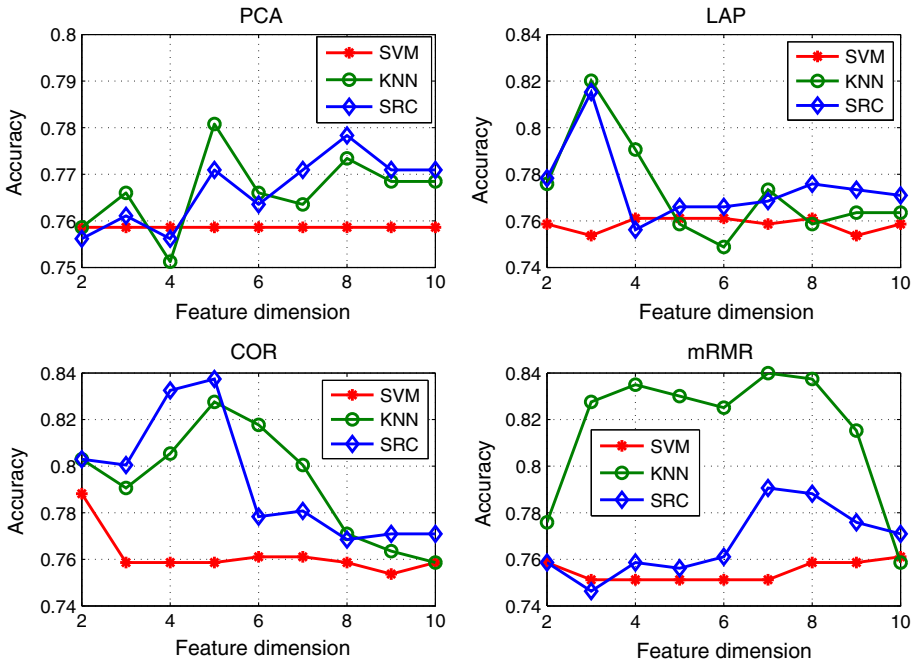


Fig. 7 Classification result for data Anneal

Table 2 Comparisons of classification accuracy (%) based on UCI data sets

Data set	PCA			LAP			COR			mRMR		
	20 %	50 %	80 %	20 %	50 %	80 %	20 %	50 %	80 %	20 %	50 %	80 %
Wine-SVM	70.79	**93.26**	92.13	75.28	86.52	91.01	47.19	83.15	**92.13**	23.60	**93.26**	**95.51**
Wine-KNN	68.54	88.76	85.39	**81.90**	87.64	88.76	**56.18**	76.40	85.39	**60.67**	92.13	93.26
Wine-SRC	**70.79**	92.58	**92.65**	80.24	**92.13**	**93.26**	53.93	**84.27**	91.01	41.57	80.90	83.15
Glass-SVM	53.27	78.50	77.57	33.64	50.47	70.09	71.03	72.90	71.03	67.29	67.29	68.22
Glass-KNN	51.40	74.77	76.64	32.71	48.60	67.29	69.16	72.90	73.83	71.96	69.16	70.09
Glass-SRC	**54.21**	**94.39**	**97.20**	**45.79**	**69.16**	**91.59**	**71.96**	**97.20**	**97.20**	**73.83**	**87.85**	**91.59**
Libras-SVM	62.22	56.67	42.78	68.33	77.78	75.56	58.89	69.44	74.44	58.33	63.33	67.22
Libras-KNN	61.72	40.56	37.78	76.11	78.33	78.89	66.11	73.89	77.78	69.44	77.22	78.89
Libras-SRC	**81.00**	**81.00**	**81.00**	**80.00**	**80.00**	**79.44**	**79.44**	**75.56**	**82.78**	**74.44**	**79.44**	**80.00**
Wquality-SVM	**49.75**	53.38	57.00	42.75	55.88	56.50	**54.75**	56.00	55.25	**54.00**	**55.13**	54.25
Wquality-KNN	44.13	48.63	52.25	**49.88**	53.25	52.13	52.63	55.75	51.13	51.00	54.88	54.38
Wquality-SRC	48.50	**55.13**	**58.75**	45.50	**58.63**	**57.38**	50.00	**59.25**	**58.13**	49.75	51.38	**56.75**
Anneal-SVM	75.86	75.86	75.86	75.37	76.11	75.86	75.86	76.11	75.86	75.86	75.12	76.11
Anneal-KNN	**75.86**	**76.60**	76.85	**82.02**	74.88	76.35	79.06	**80.05**	75.86	**77.59**	**83.99**	75.86
Anneal-SRC	75.62	76.35	**77.09**	81.53	**76.60**	**77.09**	**80.05**	78.08	**77.09**	75.86	79.06	**77.09**

LAP and COR cases. SVM is stable on this data set but the accuracy is not competitive. KNN is has advantages in mRMR case.

Table 2 is a comprehensive list of results over the UCI data sets. The highest accuracies are highlighted among SVM, KNN and SRC. The outputs of SRC have enhanced performance in most cases. Table 3 shows the standard deviation based on the accuracy from 50 to 80 %. Note the small standard deviation for SRC, highlighting the stability of SRC compared with SVM and KNN.

Experiments on Face Recognition

The extended Yale face database B [19] is used in the second experiment. In this data set, there are 2,414 faces images from 38 people, which are captured in different environments. Each face image is 54×48 pixels large. Inspired by recent work [63] using Gabor features for face recognition, the experiment is conducted to investigate SRC–SPDR framework on Gabor features. A set of Gabor filters, which contains 5 scale levels and 8 orientations, are

Table 3 Standard deviation of accuracy

Different performance	PCA	LAP	COR	mRMR
Wine-SVM	0.020	0.017	0.024	0.018
Wine-KNN	0.021	0.017	0.049	0.008
Wine-SRC	0.014	0.018	0.016	0.012
Glass-SVM	0.0047	0.11	0.0093	0.013
Glass-KNN	0.0089	0.10	0.0090	0.025
Glass-SRC	0	0.018	0	0.048
Libras-SVM	0.062	0.010	0.028	0.023
Libras-KNN	0.032	0.007	0.021	0.008
Libras-SRC	0	0.003	0.032	0.005
Wquality-SVM	0.017	0.009	0.003	0.011
Wquality-KNN	0.012	0.008	0.020	0.012
Wquality-SRC	0.008	0.002	0.017	0.008
Anneal-SVM	0	0.003	0.003	0.004
Anneal-KNN	0.004	0.006	0.019	0.038
Anneal-SRC	0.003	0.003	0.005	0.009

applied to each face image with the same parameters as in [63]. In total, there are 40 Gabor filters, and each Gaborface is with the size of 6×6. An example of Gaborfaces is shown in the left of Fig. 8. Then, the Gabor features with 1,440 dimensions are used to perform similar experiments as the one described in Sect. 5.1.

In Fig. 8, we also evaluate selecting 100–800 features by the LAP and COR methods. From this figure, one can see that the 100 and 200 features selected by these two methods are quite different. However, the accuracy of the 100 LAP and COR-selected features is similar under all three classifiers, as shown in Fig. 9. There are also obvious diversities between the selected features when the dimensions are 400 and 800. Hence, we may conclude that Gabor features have a lot of redundancy, and SPDR is necessary. In Fig. 9, SRC shows a clear performance improvement in the case of LAP and COR. Although the performance of SRC with PCA is a little worse than SVM with PCA, the accuracy is still > 92 % and remain stable at 95 % when the feature dimension is > 300.

Figure 9 shows the classification rates for number of dimensions ranging from 100 to 1,000. In particular, it is interesting to investigate the classification performance at very low dimensions. In Table 4, classification results for lower dimension (under 100) of Gabor features are listed. The number of dimensions selected varies from 10 to 100, out of the original 1,440. In the cases of COR, LAP and mRMR, SRC always obtain higher accuracy than SVM and KNN. With PCA method, the SRC accuracies are the highest when the dimension is smaller than 80.

In addition to Gabor features, we also attempted to study the performance of SRC–SPDR on the original image pixel values. The original pixel number is $54 \times 48 = 2{,}592$, which is too large for SRC and dimensionality reduction is necessary. In this experiment, three dimension reduction methods are applied on the images to reach dimensions from 10 to 100. Then SRC is applied to the dimension-reduced vector to generate the classification results in Table 5. The results are compared with results from SRC using Gabor-PCA features, SRC using random-face features [56] and SRC using downsample-face features [56]. In order to achieve the dimension from 10 to 100, the downsample process is carried out with the ratios 1/260, 1/130, 1/87, 1/65, 1/52, 1/43, 1/37, 1/32, 1/29 and 1/26,

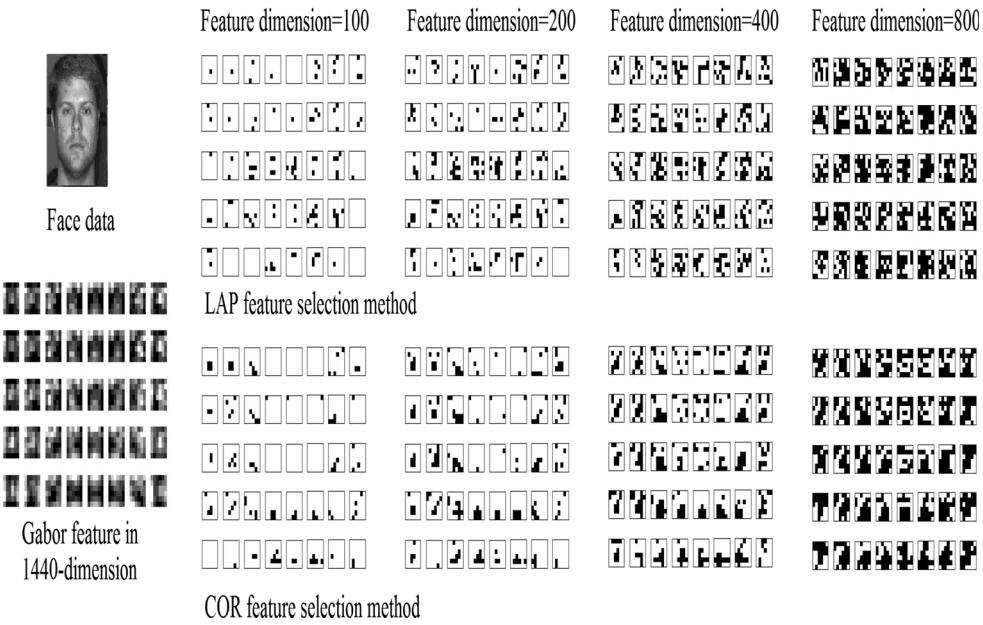


Fig. 8 Face image process, the Gabor features selected with LAP and COR are shown

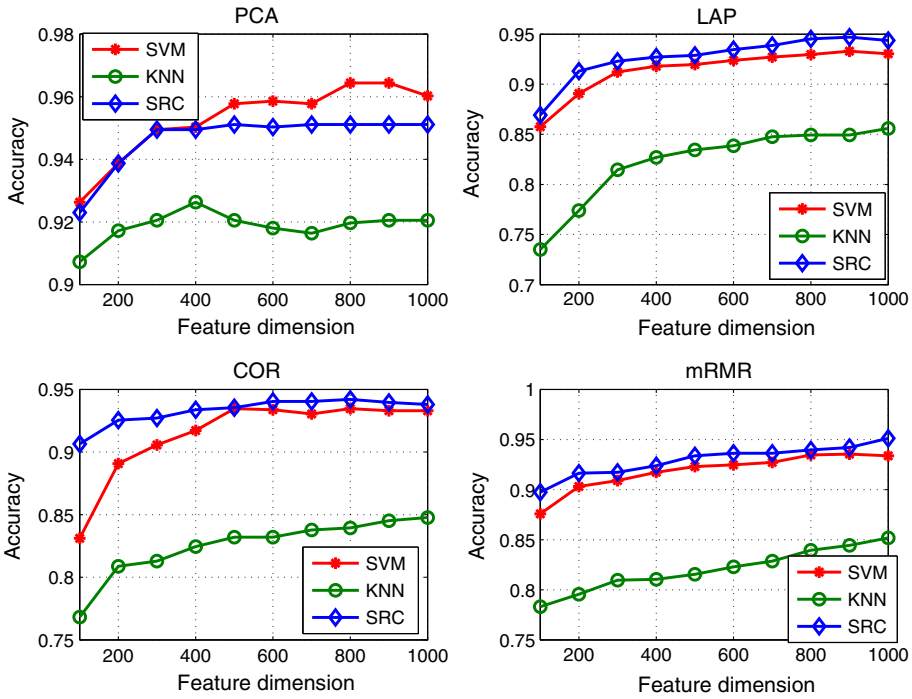


Fig. 9 Classification results for high-dimensional Gabor face features

Table 4 Classification accuracy (%) based on Gabor features

Dimension (d)	10	20	30	40	50	60	70	80	90	100
PCA-SVM	29.22	72.58	83.84	87.99	89.31	90.14	91.23	91.72	92.30	**92.63**
PCA-KNN	26.57	65.65	80.13	83.61	87.09	88.41	89.98	90.65	91.14	90.89
PCA-SRC	36.42	**73.84**	**84.35**	**88.49**	**90.23**	**91.23**	**91.31**	**92.05**	**92.30**	92.30
LAP-SVM	39.40	59.35	67.38	74.34	77.73	79.72	80.71	82.78	84.77	85.76
LAP-KNN	34.19	50.99	59.19	64.74	67.05	68.21	68.79	70.03	72.27	73.51
LAP-SRC	38.74	66.47	77.07	82.04	83.69	85.68	86.09	88.08	88.49	88.49
COR-SVM	42.96	60.51	70.20	77.90	79.72	81.54	83.53	84.69	85.60	86.75
COR-KNN	39.32	52.40	62.50	68.29	70.86	71.11	72.76	75.66	75.66	76.82
COR-SRC	42.38	64.49	77.40	83.36	85.76	88.00	89.32	89.90	89.74	90.65
mRMR-SVM	12.91	39.82	58.03	65.81	70.53	73.01	77.73	78.56	81.13	81.54
mRMR-KNN	46.27	63.41	67.05	71.61	73.18	76.57	77.57	76.41	77.07	78.39
mRMR-SRC	**48.76**	69.87	77.48	82.78	84.85	88.00	88.33	87.83	88.91	89.74

The highest accuracy among all the methods (columns) are highlighted in bold

Table 5 SRC classification accuracy (%) on different feature sets

Dimension (d)	10	20	30	40	50	60	70	80	90	100
Gabor-PCA	36.42	73.84	84.35	88.49	90.23	91.23	91.31	92.05	92.30	92.30
Pixel-PCA	48.34	**83.86**	**91.06**	**92.55**	**94.45**	**94.87**	**95.12**	**95.36**	**95.94**	**95.53**
Gabor-LAP	38.74	66.47	77.07	82.04	83.69	85.68	86.09	88.08	88.49	88.49
Pixel-LAP	43.79	71.11	80.13	83.11	85.92	87.67	87.42	89.65	90.48	91.14
Gobor-COR	42.38	64.49	77.40	83.36	85.76	88.00	89.32	89.90	89.74	90.65
Pixel-COR	**52.57**	70.61	78.06	83.69	85.26	86.26	87.75	87.91	88.49	89.16
Gobor-mRMR	48.76	69.87	77.48	82.78	84.85	88.00	88.33	87.83	88.91	89.74
Pixel-mRMR	50.66	69.95	79.06	83.86	86.42	87.25	89.24	90.31	90.07	91.31
Pixel-Random	40.23	64.07	74.59	81.21	85.35	87.25	89.16	90.56	91.97	92.63
Pixel-Downsample	38.22	63.25	79.64	82.62	81.62	89.65(*)	91.64(△)	92.64(▽)	92.80	93.29

The actual dimension for *, △ and ▽ are 61, 71 and 81 due to the downsample process

respectively. The actual dimensions are shown at the bottom of Table 5.

The highest classification rates at different dimensions are highlighted in Table 5. It is interesting to note that pixel-PCA always surpass Gabor-PCA, which indicates that SRC with dimension reduction works better in the natural signal space than in the feature space. SRC with pixel-PCA features at dimension 20 can reach 83.86 % accuracy. Compared with the results proposed in [65]with the same experiment settings as ours, their work can achieve 81.5 % in the dimension of 56. When compared with SRC on random face and SRC on downsample face [56], SRC with pixel-PCA performs significantly better, especially at very low dimensions (≤ 40). It is important to point out that the results reported in [56] were based on face images with size of 192×168, while our results are based on face images with size of 54×48.

Conclusion

A comprehensive study is conducted on a variety dimension reduction methods within the SRC framework. The purpose is to use DR techniques to improve the sparse coding process, both in efficiency and accuracy. Experiments on the UCI and face data demonstrate the effectiveness of this combination. Particularly in data Glass and data Libras Movement, SRC is able to obtain around 20 and 30 % classification accuracy improvement compared to SVM and KNN at lower dimensions. And SRC with Pixel-PCA feature can achieve more than 90 % accuracy at dimension 30 on the face data set. Based on the results, we have shown both experimentally and theoretically that SRC is efficient with dimension reduction methods. Due to the diversity of different data sets, it is not clear which dimension reduction method is the best fit for SRC, which is similar to the conclusion in the previous work [56]. However, we still could observe that PCA + SRC shows more advantages compared with other combinations, especially in the face data set.

There are many interesting future research topics along this direction. For instance, with the continuous of the big data challenge, how to integrate the sparse representation with complex data analysis tasks such as imbalanced data [27], dynamic stream data [25, 26], integrated prediction and optimization [28], among others, have become significant research topics in the society. New research foundations, principles and algorithms are needed to tackle such challenges. Furthermore, large-scale experimental studies are also needed to fully justify the effectiveness of the proposed method. Finally, as intelligent data analysis is critical in many real-world applications, how to bring the proposed techniques to a wide range of application domains is another important future research topic.

Acknowledgments This work was supported in part by the National Science Foundation under grant ECCS 1053717, Army Research Office under grant W911NF-12-1-0378, NSF-DFG Collaborative Research on "Autonomous Learning," a supplement grant to CNS 1117314, and Defense Advanced Research Projects Agency (DARPA) under grant FA8650-11-1-7152 and FA8650-11-1-7148.

References

1. Athitsos V, Sclaroff S. Boosting nearest neighbor classifiers for multiclass recognition. In: IEEE CVPR Workshops. 2005.
2. Bergeaud F, Mallat S. Matching pursuit of images. In: International conference on image processing, vol. 1. 1995. p. 53–56.
3. Bradley D, Bagnell JA. Convex coding. Tech. Report CMU-RI-TR-09-22. Pittsburgh, PA: Robotics Institute; 2009.
4. Cambria E, Hussain A. Sentic album: content-, concept-, and context-based online personal photo management system. Cognit Comput. 2012;4(4):477–96.
5. Candès EJ, Donoho DL. Curvelets: a surprisingly effective nonadaptive representation of objects with edges. In: Cohen A, Rabut C, Schumaker LL, editors. Curve and surface fitting. Saint-Malo: Vanderbilt University Press; 2000.
6. Candès EJ, Tao T. Decoding by linear programming. IEEE Trans Inf Theory. 2005;51(12):4203–15.
7. Candès and EJ, Tao T. The Dantzig selector: statistical estimation when p is much larger than n. Ann Stat. 2007;35(6):2313–51.
8. Chen X, Qi Y, Bai B, Lin Q, Carbonell JG. Sparse latent semantic analysis. In: SIAM international conference on data mining (SDM). 2011. p. 474–85.
9. Cook RD, Yin X. Dimension reduction and visualization in discriminant analysis (with discussion). N Z J Stat. 2001;43(2):147–99.
10. Cox T, Cox M. Multidimensional scaling. London: Chapman and Hall; 1994.
11. Dai JJ, Lieuand L, Rocke D. Dimension reduction for classification with gene expression microarray data. Statist Appl Genet Mol Biol. 2009;5(1):1–15.
12. Deegalla S, Boström H. Classification of microarrays with knn: comparison of dimensionality reduction methods. In: The 8th international conference on intelligent data engineering and automated, learning. 2007. p. 800–09.
13. Elad M, Aharon M. Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans Image Process. 2006;15(12):3736–45.
14. Engan K, Aase SO, Husøy JH. Multi-frame compression: theory and design. Signal Process. 2000;80(10):2121–40.
15. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B. 2008;70(5):849–911.
16. Arbib MA. The handbook of brain theory and neural networks. Cambridge, MA: MIT Press; 1995.
17. Bache K, Lichman M. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science; 2013. http://archive.ics.uci.edu/ml.
18. Gao J, Shi Q, Caetano TS. Dimensionality reduction via compressive sensing. Pattern Recognit Lett. 2012;33(9):1163–70.
19. Georghiades AS, Belhumeur PN, Kriegman DJ. From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell. 2001;23(6):643–60.
20. Gkioulekas IA, Zickler T. Dimensionality reduction using the sparse linear model. Adv Neural Inf Process Syst. 2011;24:271–9.
21. Grassi M, Cambria E, Hussain A, Piazza F. Sentic web: a new paradigm for managing social media affective information. Cognit Comput. 2011;3(3):480–9.
22. Gregor K, LeCun Y. Learning fast approximations of sparse coding. In: International conference on machine learning (Haifa, Israel). 2010. p. 399–406.

23. Gregor K, Szlam A, LeCun Y. Structured sparse coding via lateral inhibition. In: Advances in neural information processing systems (NIPS) 24. 2011.
24. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
25. He H, Chen S. Imorl: Incremental multiple objects recognition and localization. IEEE Trans. Neural Netw. 2008;19(10):1727–37.
26. He H, Chen S, Li K, Xu X. Incremental learning from stream data, IEEE Trans. Neural Netw Learn Syst. 2012;22(12):1901–14.
27. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.
28. He H, Ni Z, Fu J. A three-network architecture for on-line learning and optimization based on adaptive dynamic programming. Neurocomputing. 2012;78(1):3–13.
29. He X, Cai D, Niyogi P. Laplacian score for feature selection. In: Advances in neural information processing systems 18. Cambridge, MA: MIT Press; 2005.
30. Hu S, Yao Y, Yang Z. Mac protocol identification approach for implement smart cognitive radio. In: IEEE international conference on communications. 2012. p. 5608–12.
31. Hu S, Yao Y, Yang Z, Zheng D. Cog-prma protocol for cr users sharing a common channel with tdma primary users. In: IEEE wireless and optical communications conference. 2011. p. 1–5.
32. Huang J, Zhang T. The benefit of group sparsity. Ann Stat. 2010;38:1978–2004.
33. Kim S, Koh K, Lustig M, Boyd S, Gorinevsky D. An interior-point method for large-scale l1-regularized least squares. IEEE J Sel Top Signal Process. 2007;1(4):606–17.
34. Krause A, Cevher V. Submodular dictionary selection for sparse representation. In: International conference on machine learning (Haifa, Israel). 2010. p. 567–74.
35. La C, Do MN. Signal reconstruction using sparse tree representation. In Proceedings of Wavelets XI at SPIE Optics and Photonics, San Diego. 2005.
36. Labusch K, Barth E, Martinetz T. Sparse coding neural gas: learning of overcomplete data representations. Neurocomputing. 2009;72:1547–55.
37. Lacoste-Julien S, Sha F, Michael IJ. DiscLDA: Discriminative learning for dimensionality reduction and classification. In: Advances in neural information processing systems 22. 2008.
38. Lee H, Battle A, Raina R, Ng AY. Efficient sparse coding algorithms. In: Advances in neural information processing systems (NIPS) 19. Cambridge, MA; 2006. p. 801–8.
39. Mairal J, Bach F, Ponce J, Sapiro G. Online dictionary learning for sparse coding. In: International conference on machine learning. 2009.
40. Mallat S. A wavelet tour of signal processing. The sparse way. 3rd ed. New York: Academic Press; 2008.
41. Meinshausen N. Relaxed lasso. Comput Stat Data Anal. 2007;52:374–93.
42. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996;381(6583):607–9.
43. Palm G. Neural associative memories and sparse coding. Neural Netw. 2013;37:165–71.
44. Paschou P, Ziv E, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 2007;3(9):1–15.
45. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38.
46. Ruping S, Morik K. Support vector machines and learning about time. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4. 2003. p. 864–7.
47. Shaw B, Jebara T. Structure preserving embedding. In: The 26th annual international conference on machine learning, ICML. 2009. p. 937–44.
48. Siddiqui S, Robila S, Peng J, Wang D. Sparse representations for hyperspectral data classification. In: IEEE international geoscience and remote sensing symposium, vol. 2. 2008. p. 577–80.
49. Su L, Wang L, Chen F, Shen H, Li B, Hu D. Sparse representation of brain aging: extracting covariance patterns from structural mri. PLoS One. 2012;7(5):e6147.
50. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996;58:267–88.
51. Turk MA, Pentland AP. Face recognition using eigenfaces. In: IEEE conference on CVPR. June 1991. p. 586–91.
52. Wang C, Yan S, Zhang L, Zhang HJ. Multi-label sparse coding for automatic image annotation. In: IEEE conference on computer vision and pattern recognition (CVPR). 2009. p. 1643–50.
53. Wang J, Wang L. Sparse supervised dimension reduction in high dimensional classification. Electron J Stat. 2010;4:914–31.
54. Weinberger KQ, Packer BD, Saul LK. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In: The 10th international workshop on artificial intelligence and statistics. 2005. p. 381–8.
55. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 2000.
56. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell. 2009;31(2):210–27.
57. Xu J, He H, Man H. Dcpe co-training for classification. Neurocomputing. 2012;86:75–85.
58. Xu J, Man H. Dictionary learning based on laplacian score in sparse coding. In: Lecture notes in computer science, MLDM, vol. 6871. Springer; 2011. p. 253–64.
59. Xu J, Yang G, Man H. Sparse representation for classification with structure preserving dimension reduction. In: The 28th international conference on machine learning (ICML) workshop, (Bellevue, WA, USA). 2011.
60. Xu J, Yang G, Man H, He H. L1 graph based on sparse coding for feature selection. In: International symposium on neural networks (ISNN 2013). 2013. p. 594–601.
61. Xu J, Yin Y, Man H, He H. Feature selection based on sparse imputation. In: The international joint conference on neural networks (IJCNN). 2012. p. 1–7.
62. Yang J, Yu K, Gong Y, Huang T. Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on CVPR. 2009. p. 1794–801.
63. Yang M, Zhang L. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In: Computer Vision–ECCV 2010. Berlin, Heidelberg: Springer; 2010. p. 448–61.
64. Zeng X, Luo S, Li Q. An associative sparse coding neural network and applications. Neurocomputing. 2010;73:684–9.
65. Zhang L, Yang M, Feng Z, Zhang D. On the dimensionality reduction for sparse representation based face recognition. In: The 20th international conference on pattern recognition (ICPR). 2010. p. 1237–40.
66. Zhang L, Zhu P, Hu Q, Zhang D. A linear subspace learning approach via sparse coding. In: IEEE international conference on computer vision. 2011. p. 755–61.
67. Zheng C, Huang DS, Shang L. Feature selection in independent component subspace for microarray data classification. Neurocomputing. 2006;69(16–18):2407–10.
68. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc B. 2005;67(4):301–20.
69. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. J Comput Graph Stat. 2004;15:265–86.
70. Zylberberg J, Murphy JT, DeWeese MR. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields. PLoS Comput Biol. 2011;7(10).